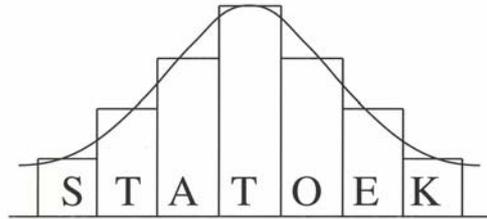


# Institut für Statistik und Ökonometrie



## **Zählmodellen (Count Data Models) Ansätze und Anwendungen**

Verena Dexheimer

Arbeitspapier Nr. 23 (Mai 2002)

**Johannes Gutenberg-Universität  
Fachbereich Rechts-  
und Wirtschaftswissenschaften**

Haus Recht und Wirtschaft II  
D 55099 Mainz  
Tel: 0 61 31 - 39 22551      Fax: 0 61 31 - 39 23717  
E-Mail: [STATOEK@Uni-Mainz.de](mailto:STATOEK@Uni-Mainz.de)

**Herausgeber: Univ.-Prof. Dr. Peter M. Schulze**

ISSN Nr. 1430 - 2136

Institut für Statistik und Ökonometrie  
Johannes Gutenberg-Universität Mainz  
Fachbereich Rechts- und Wirtschafts-  
wissenschaften  
Haus Recht und Wirtschaft II

D 55099 Mainz

Herausgeber: Univ.-Prof. Dr. P.M. Schulze

© 2002 Institut für Statistik und Ökonometrie, Mainz  
ISSN Nr. 1430-2136

## Gliederung

1	Einleitung .....	2
2	Theorie von Zählmodellen.....	2
	2.1 Das Poisson-Regressionsmodell.....	2
	2.2 Das Regressionsmodell der Negativen Binomialverteilung.....	4
	2.3 Güte der Schätzungen.....	6
	2.4 Spezifikationstests .....	8
3	Empirische Anwendung auf Patentdaten.....	11
	3.1 Datensatz und Datenberechnungen .....	11
	3.2 Deskriptive Statistik .....	12
	3.3 Spezifikation der Zählansätze.....	13
	3.4 Schätzung und Beurteilung der Zählansätze.....	14
	3.5 Vergleich der verschiedenen Schätzergebnisse.....	20
4	Zusammenfassung und Ausblick.....	21
	Literatur .....	24

## Zusammenfassung

Die Arbeit umfaßt die Darstellung und Anwendung von Regressionsmodellen für Zählmodellen, deren abhängige Variable nur nichtnegative ganzzahlige Werte annehmen kann. Die Spezifikationen des Poisson- und des NegBin 2-Modells mit den zugehörigen Schätzverfahren und Beurteilungskriterien werden für den Querschnittsdaten-Bereich vorgestellt. Als Entscheidungshilfe für das eine oder andere Modell werden Tests der Daten auf Overdispersion herangezogen. Im empirischen Teil der Arbeit bestätigt sich der in der Literatur dargestellte stark positive Zusammenhang zwischen den Patentanmeldungen und den FuE-Beschäftigten bzw. den FuE-Ausgaben. Dabei wird angesichts zu großer Streuung der Daten das NegBin 2-Modell dem Poisson-Modell vorgezogen.

## Summary

This paper deals with the description and application of regression models for count data where the dependent variable takes only nonnegative integer values. The specification of the Poisson- and the NegBin 2-Model with their estimation procedures and goodness-of-fit-statistics are presented for cross-section data. To come to a decision between the two models tests for overdispersed data are used. The strong positive relationship between patent applications and the R&D personnel respectively R&D expenditures which is presented in the literature could be confirmed with this application. Because of overdispersion in the data the NegBin2-Model can be preferred to the Poisson-Model.

## 1 Einleitung

Regressionsmodelle, deren abhängige Variable die Anzahl von auftretenden Ereignissen in einem gegebenen Zeitintervall abbildet, werden in der Literatur als Zähldatenmodelle bezeichnet. Die Zählvariable nimmt dabei nur nichtnegative Werte aus der Menge der natürlichen Zahlen an und wird durch einen Vektor von unabhängigen Variablen beschrieben (Winter, 2001, S. 30).

Nach einer theoretischen Einführung wird im empirischen Teil mit Hilfe dieser Zähldatenmodelle geprüft, wie die Zahl der angemeldeten Patente von FuE-Inputgrößen wie FuE-Ausgaben bzw. FuE-Beschäftigtenzahlen abhängt.

## 2 Theorie von Zähldatenmodellen

Hier werden nur die theoretischen Grundlagen von Zähldatenmodellen vorgestellt, die für die spätere Anwendung auf Patentdaten nötig sind.

### 2.1 Das Poisson-Regressionsmodell

Das Basismodell für Zählvariablen im Querschnittsdaten-Bereich bildet das Poisson-Regressionsmodell (Cameron/ Trivedi, 1998, S. 9). Dabei gibt die Zufallsvariable  $y_i$  die Zahl der Ereignisse für einen von insgesamt  $n$  unabhängigen Beobachtungswerten in einem bestimmten Zeitintervall an. Der erwartete Wert der Zählvariablen  $y_i$  hängt vom Vektor der erklärenden Variablen  $\mathbf{x}_i$  und dem Vektor der Schätzparameter  $\boldsymbol{\beta}$  ab. Im folgenden wird der erwartete Wert als konditionaler Mittelwert  $\mu$  bezeichnet (Cameron/ Trivedi, 1998, S. 9):

$$(1) \quad E[y_i | \mathbf{x}_i] = \mu(\mathbf{x}_i, \boldsymbol{\beta}), \quad i = 0, 1, 2, \dots, n.$$

Dabei wird  $\mu$  wie folgt spezifiziert:

$$(2) \quad \mu_i = \exp(\mathbf{x}_i' \boldsymbol{\beta}),$$

um zu gewährleisten, daß  $\mu_i > 0$  und die erwarteten Werte der Zählvariablen damit nicht negativ werden (Ronning, 1991, S. 163).

Wenn die Ereignisse einem Poisson-Prozeß folgen und daher proportional zur Länge des

Zeitintervalls auftreten (Ronning, 1991, S. 159), bestimmt sich die bedingte Wahrscheinlichkeit für die Zählvariable  $y_i$  nach der Wahrscheinlichkeitsdichtefunktion der Poissonverteilung (Cameron/ Trivedi, 1998, S. 9):

$$(3) \quad f(y_i | \mathbf{x}_i) = \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!}; \quad y_i = 0, 1, 2, \dots; \quad i = 0, 1, 2, \dots, n.$$

Da es sich bei der Poissonverteilung um eine Ein-Parameter-Verteilung handelt, sind Mittelwert und Varianz von  $y_i$  annahmegemäß gleichgroß. Es gilt (Winkelmann, 2000, S. 66):

$$(4) \quad \mu_i = \exp(\mathbf{x}_i' \boldsymbol{\beta}) = E[y_i | \mathbf{x}_i] = V[y_i | \mathbf{x}_i].$$

Diese Eigenschaft wird als „Equidispersion“<sup>1</sup> (Cameron/ Trivedi, 1998, S. 4) bezeichnet. Zu deren Erfüllung muß die Wahrscheinlichkeit für das Auftreten eines Ereignisses über die Zeit konstant sein (Homogenitätsannahme) und sich unabhängig vom Eintreten anderer Ereignisse bestimmen (Unabhängigkeitsannahme) (King, 1989, S. 764f). Die Varianz des Poisson-Modells ist demnach keine Konstante und das Modell ist bereits heteroskedastisch in sich (Cameron/ Trivedi, 1998, S. 10).

Da das Poisson-Modell den nichtlinearen Regressionsmodellen zuzuordnen ist, wird die Maximum-Likelihood-Schätzmethode (MLS) verwendet. Die Log-Likelihood-Funktion sieht wie folgt aus (Cameron/ Trivedi, 1998, S. 61f):

$$(5) \quad \mathcal{L}(\boldsymbol{\beta}) = \ln L(\boldsymbol{\beta}) = \sum_{i=1}^n \{y_i \mathbf{x}_i' \boldsymbol{\beta} - \exp(\mathbf{x}_i' \boldsymbol{\beta}) - \ln y_i!\}.$$

Mit Hilfe der ML-Schätzung lassen sich konsistente und asymptotisch effiziente Schätzwerte für den Vektor  $\hat{\boldsymbol{\beta}}_p$  bestimmen, wenn die spezifizierte Verteilung der wahren Verteilung der Zählvariablen entspricht (Winkelmann, 2000, S. 79). Wird ein Poisson-Modell geschätzt, muß demzufolge die wahre Wahrscheinlichkeitsdichtefunktion mit der Dichte der Poissonverteilung übereinstimmen. Daraus resultiert, daß der konditionale Mittelwert und die konditionale Varianz gleich groß sein müssen (vgl. Gleichung (4)).

---

<sup>1</sup>Dieser englische Begriff wird verwendet, da in der deutschsprachigen Literatur keine geeignete Übersetzung existiert.

In vielen Anwendungen wird die Restriktion der Equidispersion nicht erfüllt. Meist tritt das Problem der sogenannten „Overdispersion“<sup>2</sup> auf, wobei die Varianz den Mittelwert übersteigt (Cameron/ Trivedi, 1986, S. 31), die Streuung demnach zu groß ist. Ein Grund dafür kann in der Verletzung der Unabhängigkeitsannahme liegen (King, 1989, S. 765). Der Kauf eines Produktes kann zum Beispiel bedingt sein durch den vorangegangenen Kauf desselben Produkts im gleichen Zeitintervall. Oft wird auch die Wahrscheinlichkeit für das Auftreten des Ereignisses über die Zeit variieren und damit die Homogenitätsannahme verletzt. Dies ist etwa bei saisonalen Effekten der Fall (Winter, 2001, S. 31).

Führt man im Falle zu großer Streuung die Poisson-ML-Schätzung durch, erhält man unter der Bedingung des korrekt spezifizierten konditionalen Mittelwertes zwar weiterhin konsistente Schätzwerte  $\hat{\beta}_p$ , jedoch sind diese nicht mehr effizient und deren Standardfehler sind verzerrt. Die darauf basierenden Hypothesentests sind folglich ungültig (Cameron/ Trivedi, 1996, S. 374).

Zur Behebung dieses Nachteils spezifizieren alternative Ansätze anstelle der Poissonverteilung eine allgemeinere Verteilung, die nicht die Restriktion der Equidispersion beinhaltet. Am häufigsten wird die Negative Binomialverteilung benutzt, die die Varianz als bestimmte Funktion des Mittelwertes spezifiziert. Standard-ML-Schätzer dieser Modelle sind bei Overdispersion effizienter als Poisson-ML-Schätzer (Cameron/ Trivedi, 1996, S. 374).

## 2.2 Das Regressionsmodell der Negativen Binomialverteilung

Ein Grund für zu große Streuung in den Daten kann in unbeobachteter Heterogenität der Beobachtungswerte liegen. Um dies zu berücksichtigen, wird der konditionale Poisson-Mittelwert im NegBin-Modell um einen stochastischen Fehlerterm  $\nu_i$  multiplikativ erweitert (Cameron/ Trivedi, 1998, S. 98ff):

$$(6) \quad E[y_i | \mu_i, \nu_i] = \mu_i^* = \mu_i \nu_i.$$

Der Fehlerterm  $\nu_i$  umfaßt neben der Heterogenität zusätzlich systematische Spezifikationsfehler des Mittelwertes wie bspw. unbeobachtete erklärende Variablen (Greene, 2000, S. 886) und wird meist als iid angenommen. Als Wahrscheinlichkeitsdichtefunktion des

---

<sup>2</sup> Auch hier wird der englische Begriff verwendet, da in der deutschsprachigen Literatur keine zweckdienliche Übersetzung existiert.

Fehlerterms  $v_i$  wird im NegBin-Modell eine Gammaverteilung mit den Parametern  $E[v] = 1$  und  $V[v] = \alpha$  ausgewählt. Der erwartete Wert von  $y_i$  unter der Bedingung von  $\mathbf{x}_i$  und  $v_i$  ist weiterhin poissonverteilt.

Für die Wahrscheinlichkeitsdichte der NegBin-Verteilung mit  $\alpha$  als „Overdispersionsparameter“ erhält man (Cameron/ Trivedi, 1998, S. 98ff):

$$(7) \quad f(y_i | \mathbf{x}_i) = \frac{\Gamma(y_i + \alpha^{-1})}{\Gamma(y_i + 1) \Gamma(\alpha^{-1})} \left( \frac{\mu_i}{\mu_i + \alpha^{-1}} \right)^{y_i} \left( \frac{\alpha^{-1}}{\mu_i + \alpha^{-1}} \right)^{\alpha^{-1}}.$$

Hierbei handelt es sich um das sogenannte NegBin 2-Modell mit den ersten zwei konditionalen Momenten:

$$(8) \quad E[y_i | \mu_i, \alpha] = \mu_i = \exp(\mathbf{x}_i' \boldsymbol{\beta});$$

$$(9) \quad V[y_i | \mu_i, \alpha] = \mu_i + \alpha \mu_i^2 = \exp(\mathbf{x}_i' \boldsymbol{\beta}) + \alpha [\exp(\mathbf{x}_i' \boldsymbol{\beta})]^2.$$

Falls der Overdispersionsparameter  $\alpha = 0$  ist, reduziert sich das NegBin-Modell zu einem Poisson-Modell (Cameron/ Trivedi, 1998, S. 75).

Bei der NegBin 2-ML-Schätzung wird neben den ML-Schätzern der Regressionsparameter  $\hat{\boldsymbol{\beta}}_{NB2}$  zusätzlich ein Overdispersionsparameter  $\hat{\alpha}_{NB2}$  bestimmt. Die verwendete Log-Likelihood-Funktion unterscheidet sich von der Log-Likelihood-Funktion des Poisson-Modells und sieht wie folgt aus (Cameron/ Trivedi, 1998, S. 71):

$$(10) \quad \ln L(\alpha, \boldsymbol{\beta}) = \sum_{i=1}^n \left\{ \left( \sum_{j=0}^{y_i-1} \ln(j + \alpha^{-1}) \right) - \ln y_i! \right. \\ \left. - (y_i + \alpha^{-1}) \cdot \ln[1 + \alpha \exp(\mathbf{x}_i' \boldsymbol{\beta})] + y_i \ln \alpha + y_i \mathbf{x}_i' \boldsymbol{\beta} \right\}.$$

Bei negativ binomialverteilten Daten erhält man NegBin 2-ML-Schätzer, die relativ effizienter als die Poisson-ML-Schätzer sind (Cameron/ Trivedi, 1996, S. 374). Die  $\hat{\boldsymbol{\beta}}_{NB2\_ML}$ , deren Standardfehler und  $\hat{\alpha}_{NB2\_ML}$  können dann konsistent geschätzt werden. Da die ML-Methode keine Verteilungsmißspezifikation zuläßt, kommt es bei nicht negativ binomialverteilter Zählvariable zu inkonsistenten Standardfehlerschätzungen aufgrund von inkonsistenten Schätzwerten für  $\hat{\alpha}_{NB2\_ML}$  (Cameron/Trivedi, 1998, S. 74f).

In einem solchen Fall muß eine andere Schätzmethode verwendet werden, die nicht mehr voraussetzt, daß die spezifizierte Wahrscheinlichkeitsverteilung der Daten notwendigerweise die wahre ist. Eine solche Schätzmethode stellt die Quasi-Generalized-Pseudo-ML-Schätzung (QGPMML) dar. Um korrekte Standardfehler zu schätzen müssen hier lediglich der konditionale Mittelwert und die konditionale Varianz von  $y_i$  den Daten entsprechend spezifiziert sein (Cameron/ Trivedi, 1986, S. 35).

Dabei wird folgende Varianzfunktion zur Schätzung der  $\beta$  verwendet (Cameron/ Trivedi, 1998, S. 63):

$$(11) \quad \omega_i = \omega(\mu_i, \alpha) = (\mu_i + \alpha\mu_i^p).$$

Die zu maximierende Log-Likelihood-Funktion basiert auf einer Funktion, die einen Overdispersionsparameter  $\alpha$  beinhaltet, der zunächst nach der Gleichung (12) konsistent bestimmt werden muß:

$$(12) \quad \tilde{\alpha}_{QGPMML\_NB2} = \frac{1}{n-k} \sum_{i=1}^n \frac{\{(y_i - \hat{\mu}_i)^2 - \hat{\mu}_i\}}{\hat{\mu}_i^2}.$$

$\tilde{\alpha}$  wird dann bei der QGPMML von  $\beta$  benutzt (Dionne/ Vanasse, 1992, S. 155). Dieses zweistufige Verfahren ermittelt konsistente Schätzwerte  $\hat{\beta}_p$ , wenn Varianz und Mittelwert korrekt spezifiziert sind (Cameron/ Trivedi, 1998, S. 33).

Im Fall der NegBin 2-QGPMML mit Mittelwert  $\mu_i$  und Varianz  $\mu_i + \alpha\mu_i^2$  sieht die Log-Likelihood-Funktion wie folgt aus (Cameron/ Trivedi, 1998, S. 73):

$$(13) \quad \ln L(\tilde{\alpha}, \beta) = \sum_{i=1}^n \left\{ -\tilde{\alpha}^{-1} \ln(1 + \tilde{\alpha}\mu_i) + y_i \ln\left(\frac{\tilde{\alpha}\mu_i}{1 + \tilde{\alpha}\mu_i}\right) + b(y_i, \tilde{\alpha}) \right\}.$$

Liegt jedoch keine Mißspezifikation der Daten vor, ist die MLS vorzuziehen, da sie relativ effizienter ist als die QGPMML (Cameron/ Trivedi, 1986, S. 35).

### 2.3 Güte der Schätzungen

Im Bereich der Zählmodellen werden in der Literatur eine Vielzahl von Pseudo-R<sup>2</sup>-Meßgrößen zur Beurteilung der Leistungsfähigkeit des geschätzten Modells dargestellt. Diese beruhen nicht wie das R<sup>2</sup> des klassischen Regressionsmodells auf der Summe von

quadrierten, ungewichteten Residuen, sondern bauen auf gewichteten Restwerten auf. Damit wird die modellinhärente Heteroskedastie von Zähldatenmodellen berücksichtigt. Je nach vorliegender Residuendefinition unterscheidet man verschiedene  $R^2$ -Maße (vgl. im einzelnen Cameron/ Windmeijer, 1996, S. 209-220).

### 2.3.1 $R^2$ -Meßgrößen auf Basis von Pearson-Residuen

Durch Gewichtung der Residuen  $r_i = (y_i - \hat{\mu}_i)$  mit der Wurzel des Varianz-Schätzwertes  $\hat{\omega}_i$  von  $y_i$  erhält man quadriert und aufsummiert die sogenannte Pearson-Statistik:

$$(14) \quad P = \sum_{i=1}^n \frac{(y_i - \hat{\mu}_i)^2}{\hat{\omega}_i}.$$

Die folgende Größe  $R_P^2$  basiert auf dem Vergleich der Pearson-Statistiken von unrestringiertem und restringiertem Modell.

Im Poisson-Fall folgt mit  $\omega = \mu$ :

$$(15) \quad R_{P,P}^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{\mu}_i)^2 / \hat{\mu}_i}{\sum_{i=1}^n (y_i - \bar{y}_k)^2 / \bar{y}_k}.$$

$\bar{y}_k$  entspricht dem geschätzten Mittelwert des Konstantenmodells.

Nachteilig ist, daß  $R_{P,P}^2$  den Wert 1 nicht erreichen kann und sich unter Umständen verringert, wenn weitere Regressoren aufgenommen werden.  $R_{P,P}^2$  kann sogar negative Werte annehmen.

Bei einem NegBin 2-Modell mit  $\omega = \mu + \alpha\mu^2$  folgt:

$$(16) \quad R_{P,NB2}^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{\mu}_i)^2 / (\hat{\mu}_i + \alpha\hat{\mu}_i^2)}{\sum_{i=1}^n (y_i - \bar{y}_k)^2 / (\bar{y}_k + \alpha\bar{y}_k^2)}.$$

Auch  $R_{P,NB2}^2$  kann negativ werden und abnehmen, wenn weitere Regressoren aufgenommen werden (Cameron/ Trivedi, 1998, S. 155).

### 2.3.2 $R^2$ -Meßgrößen auf Basis von Deviance-Residuen

Aufgrund der Nachteile der oben genannten  $R^2$ -Größen wird nun eine alternative Definition der Restwerte betrachtet.

Wenn die Wahrscheinlichkeitsverteilung der Variablen  $y_i$  der Linearen Exponential Familie (LEF) angehört, kann man folgende Deviance-Residuen verwenden (Cameron/ Trivedi, 1998, S. 141):

$$(17) \quad d_i = \text{sign}(y_i - \hat{\mu}_i) \sqrt{2\{\mathcal{L}(y_i) - \mathcal{L}(\hat{\mu}_i)\}}.$$

Die Deviance entspricht der zweifachen Differenz zwischen der maximal erreichbaren Log-Likelihood-Funktion  $\mathcal{L}(y)$  und der Log-Likelihood-Funktion des geschätzten Modells  $\mathcal{L}(\hat{\mu})$ .<sup>3</sup>

Bildet man den Quotienten aus der Deviance des Poisson-Modells und des Konstantenmodells und zieht diesen von 1 ab, erhält man folgendes  $R_{DEV,P}^2$  (Greene, 1998, S. 601):

$$(18) \quad R_{DEV,P}^2 = 1 - \frac{\sum_{i=1}^n \{y_i \ln(y_i / \hat{\mu}_i) - (y_i - \hat{\mu}_i)\}}{\sum_{i=1}^n y_i \ln(y_i / \bar{y}_k)}.$$

$R_{DEV,P}^2$  liegt zwischen 0 und 1 und wird bei Hinzunahme von Regressoren nicht kleiner.

Für das NegBin 2-Modell ergibt sich folgendes  $R_{DEV,NB2\_ML}^2$ :

$$(19) \quad R_{DEV,NB2\_ML}^2 = 1 - \frac{\sum_{i=1}^n \left\{ y_i \ln\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i + \alpha^{-1}) \ln\left(\frac{y_i + \alpha^{-1}}{\hat{\mu}_i + \alpha^{-1}}\right) \right\}}{\sum_{i=1}^n \left\{ y_i \ln\left(\frac{y_i}{\bar{y}_k}\right) - (y_i + \alpha^{-1}) \ln\left(\frac{y_i + \alpha^{-1}}{\bar{y}_k + \alpha^{-1}}\right) \right\}}.$$

$R_{DEV,NB2}^2$  liegt ebenfalls zwischen 0 und 1, bleibt jedoch bei Hinzunahme von Regressoren nicht unbedingt unverändert.

## 2.4 Spezifikationstests

Einen weiteren wichtigen Bestandteil der Modellbeurteilung bilden die Spezifikationstests. Um Aussagen über den Einfluß der aufgenommenen Regressoren auf die Zählvariable machen zu können, werden die Regressionsparameter auf Signifikanz getestet.

---

<sup>3</sup> Man bezeichnet die Deviance-Statistik auch als  $G^2$ -Statistik, die Null wird, wenn die Anpassung perfekt ist (Greene, 2000, S. 883).

Zusätzlich untersucht man, ob die angenommene Modellart den Daten gerecht wird. Mit Hilfe von Tests, die die Daten auf vorliegende Overdispersion prüfen, kann eine Auswahl zwischen dem grundlegenden Poisson-Modell und dem allgemeineren NegBin-Modell getroffen werden. Die Nullhypothese dieser sogenannten Overdispersionstests besagt, daß die Zählvariable  $y_i$  poissonverteilt ist mit (Cameron/ Trivedi, 1986, S. 40ff):

$$(20) \quad E[y_i | \mathbf{x}_i, \boldsymbol{\beta}] = V[y_i | \mathbf{x}_i, \boldsymbol{\beta}] = \mu_i.$$

In der Alternativhypothese wird für die Varianz eine allgemeinere Funktion zugelassen:

$$(21) \quad V[y_i | \mathbf{x}_i, \boldsymbol{\beta}] = \mu_i + \alpha \mu_i^p \quad (\text{z.B. für } p = 2 \text{ folgt die Negbin 2-Varianzfunktion}).$$

Ein Signifikanztest auf  $\alpha = 0$  prüft auf Overdispersion. Ist die Nullhypothese gültig, kann ein Poisson-Modell angewendet werden. Im Falle von  $\alpha > 0$  liegt Overdispersion vor, bei der Anwendung eines Poisson-Modells würde man daher einen Spezifikationsfehler begehen. Die Spezifikation eines Negbin 2-Modells kann dann zweckmäßig sein.

#### 2.4.1 Likelihood-Ratio-Test (LR-Test)

Die LR-Teststatistik bestimmt sich aus der Differenz der geschätzten Log-Likelihood-Funktionen des unrestringierten ( $\hat{\mathcal{L}}_u$ ) und des restringierten ( $\hat{\mathcal{L}}_r$ ) Modells. Die folgende LR-Teststatistik ist  $\chi^2$ -verteilt mit  $k$  Freiheitsgraden, wobei  $k$  der Zahl der Restriktionen entspricht (z.B. Winkelmann, 2000, S. 104f):

$$(22) \quad LR = -2(\hat{\mathcal{L}}_r - \hat{\mathcal{L}}_u) \sim \chi_k^2 \quad (\text{rechtsseitiger Test}).$$

Beim Overdispersionstest bestimmt sich  $\hat{\mathcal{L}}_r$  aus der Poisson-Schätzung und  $\hat{\mathcal{L}}_u$  aus der NegBin-Schätzung, somit ist  $k = 1$ . Der Nachteil dieses Tests besteht darin, daß man beide Modelle schätzen muß, um eine Aussage über die Overdispersion treffen zu können.

Bei allgemeinen Signifikanztests ergibt sich  $k$  aus den Anzahl der zusätzlich aufgenommenen Regressoren im unrestringierten Modell. Der LR-Test wird außerdem häufig benutzt, um die Signifikanz aller Parameter gemeinsam zu testen, was dem F-Test im klassischen linearen Regressionsmodell entspricht. Dabei wird die Log-Likelihood-Funktion des geschätzten (unrestringierten) Modells mit der des (restringierten) Modells verglichen, das nur eine Konstante enthält.

### 2.4.2 Wald-Test

Für die Wald-Teststatistik muß lediglich das unrestringierte Modell geschätzt werden (z.B. Winkelmann, 2000, S. 106).

Der Test prüft mit folgender „t“-Statistik<sup>4</sup> die Signifikanz des jeweiligen Parameters  $\theta$  :

$$(23) \quad Wald(= "t") = \frac{\hat{\theta}}{\sqrt{Var(\hat{\theta})}} \sim N(0,1).^5$$

Beim Overdispersionstest handelt sich um einen rechtsseitigen Signifikanztest des Overdispersionsparameters  $\alpha$  mit standardnormalverteilter Z-Prüfgröße, da nur auf  $H_1 : \alpha > 0$  getestet wird. Testet man auf Signifikanz des Parameters  $\beta$  mit  $H_1 : \beta \neq 0$ , ist die zweiseitige, standardnormalverteilte Z-Prüfgröße zu benutzen (Cameron/ Trivedi, 1998, S. 47).

### 2.4.3 Lagrange-Multiplier-Test (LM-Test)<sup>6</sup>

Die LM-Teststatistik des Overdispersionstest läßt sich bei der Prüfung zwischen Poisson- und NegBin-Modell auf folgende  $\chi^2$ -verteilte Teststatistik mit  $k = 1$  Freiheitsgraden vereinfachen (Greene, 2000, S. 886):

$$(24) \quad LM = \left[ \frac{\sum_{i=1}^n [(y_i - \hat{\mu}_i)^2 - y_i]}{\sqrt{2 \sum_{i=1}^n \hat{\mu}_i^2}} \right]^2 \sim \chi_k^2 \quad (\text{rechtsseitiger Test}).$$

Ein Vorteil des LM-Tests auf Overdispersion liegt darin, daß nur das Poisson-Modell geschätzt werden muß.

### 2.4.4 Overdispersionstest auf Basis einer KQ-Regression<sup>7</sup>

Ein weiterer Test auf Overdispersion (nach Cameron/ Trivedi, 1990, S. 347-364) trifft keine Behauptungen über die Verteilung der Daten, sondern spezifiziert lediglich die beiden

<sup>4</sup> Die Bezeichnung erfolgt in Analogie zum t-Test im Standard-Regressionsmodell.

<sup>5</sup> Die t-Statistik entspricht der Wurzel aus der allgemeinen Wald- $\chi^2$ -Statistik. Zieht man die Wurzel aus den kritischen  $\chi^2(1)$ -Werten entsprechen die Werte denen der Normalverteilung (Cameron/Trivedi, 1998, S.47).

<sup>6</sup> In der Literatur wird dieser Test auch als „Score-Test“ bezeichnet.

<sup>7</sup> Dieser Test heißt in der englischsprachigen Literatur „Regression-Based Test for Overdispersion“.

ersten Momente. Dabei liegt die Annahme zugrunde, daß zwischen der Varianz und dem Mittelwert eine Regressionsbeziehung besteht.

Die Hypothesen sehen demnach wie folgt aus:

$$(25) \quad H_0 : \quad V[y_i | \mathbf{x}_i, \boldsymbol{\beta}] = \mu_i;$$

$$(26) \quad H_1 : \quad V[y_i | \mathbf{x}_i, \boldsymbol{\beta}] = \mu_i + \alpha g(\mu_i).$$

Formt man die Alternativhypothese um, ergibt sich folgende KQ-Regression mit den geschätzten  $\hat{\mu}_i$  des Poisson-Modells [mit  $g(\mu_i) = \mu_i$  bzw.  $g(\mu_i) = \mu_i^2$ ]:

$$(27) \quad [(y_i - \hat{\mu}_i)^2 - y_i] = \alpha g(\hat{\mu}_i) + \varepsilon_i.$$

Der Parameter  $\alpha$  wird mit Hilfe einer Weighted-Least-Squares-Schätzung bestimmt, da die Fehlerterme  $\varepsilon_i$  heteroskedastisch sind. Beim sogenannten optimalen regressionsbasierten Test werden folgende Gewichte gewählt (Cameron/ Trivedi, 1990, S. 353):

$$(28) \quad w_i = \hat{\mu}_i \sqrt{2}.$$

Ein t-Test auf  $\alpha = 0$  prüft die Daten dabei auf zu große Streuung. Ist der Parameter signifikant, liegt Overdispersion vor.

### 3 Empirische Anwendung auf Patentdaten

Nach der theoretischen Darstellung von Zähldatenmodellen im Abschnitt 2 erfolgt nun deren Anwendung auf die Fragestellung, wie die Zahl der Patentanmeldungen von der Anzahl an FuE-Beschäftigten und der Höhe der internen FuE-Ausgaben<sup>8</sup> im „Wirtschaftssektor“<sup>9</sup> abhängt.

#### 3.1 Datensatz und Datenberechnungen

Die Datenbasis zur Analyse dieser Abhängigkeit umfaßt zunächst die Zahl der Patentanmeldungen des „Wirtschaftssektors“ nach Bundesländern. Diese ist für das Jahr 1998 aus der Summe der Patentanmeldungen pro Bundesland (Greif, 2001, S. 143) und den prozentualen Anteilen, die davon auf den „Wirtschaftssektor“ entfallen (Greif, 2001, S. 146), berechnet worden.

<sup>8</sup> Ausgaben für eigens durchgeführte FuE durch die Unternehmen.

<sup>9</sup> „Danach sind dem Wirtschaftssektor Unternehmen, Organisationen und Institutionen zuzurechnen, deren Haupttätigkeit es ist, Güter und Dienstleistungen zu produzieren, um sie am Markt zu einem Preis anzubieten, der annähernd wenigstens die Herstellungskosten deckt.“ (Wissenschaftsstatistik GmbH, 2001, S. 36).

Ferner gehen die Zahl der FuE-Beschäftigten sowie die internen FuE-Ausgaben des „Wirtschaftssektors“ als unabhängige Größen in die Betrachtung mit ein. Diese Angaben sind aus dem „FuE-Datenreport 2001“ des Stifterverbandes für die Deutsche Wissenschaft (S. 32-33) entnommen.

Im Zuge der „FuE-Erhebung 1999“ wurden hierzu Unternehmen und Institutionen für Gemeinschaftsforschung<sup>10</sup> über Ist-Daten und Prognosen für kommende Jahre befragt. Die beiden Inputgrößen stammen aus dem Jahr 1997, um zu berücksichtigen, daß es aufgrund von FuE-Aufwendungen erst nach einer gewissen zeitlichen Verschiebung zu einer Patententwicklung kommt. Eine Untersuchung für Deutschland ermittelte für eine solche Verschiebung einen Zeitraum von 1 bis 2 Jahren (Greif, 1995, S. 229 ff). Die Daten finden sich in Tabelle 3.8 auf S. 23.

Bei der Beurteilung der Ergebnisse ist zu beachten, daß die verwendeten Variablen aus verschiedenen Datenquellen stammen. Nach Greif werden dem Bereich „Wirtschaft“ alle Patente zugerechnet, die nicht auf Freie Erfinder und die Wissenschaft entfallen (Greif, 2001, S. 146). Im FuE-Datenreport wird genauer spezifiziert, welche Institutionen dem Wirtschaftssektor zuzuordnen sind.

### 3.2 Deskriptive Statistik

Erste Untersuchungen des Datensatzes ergaben, daß im Jahr 1997, bei einer Standardabweichung von 22.090,77 Personen, pro Bundesland im Mittel 17.891,88 Personen in FuE für den Wirtschaftssektor tätig waren. Die Unternehmen gaben hierbei durchschnittlich 3.533,75 für FuE aus, wobei die Standardabweichung bei 4.704,76 Millionen DM lag. Im Mittel über alle Bundesländer wurden in der Wirtschaft im darauffolgenden Jahr 2.409,125 Patente angemeldet. Die Standardabweichung war auch hier mit 3.342,96 Patenten sehr hoch, was einer Varianz von 11.175.360,8 entspricht.

Die hohen Standardabweichungen können zum einen aus der geringen Zahl von Beobachtungswerten resultieren, zum anderen kann ein Vergleich von Stichprobenvarianz mit dem Stichprobenmittelwert der Patentvariablen bereits einen Anhaltspunkt auf vorhandene Overdispersion geben (Cameron/ Trivedi, 1998, S. 77).

Da die Varianz der Patente den Mittelwert um ein Vielfaches übersteigt, muß damit gerechnet werden, daß die Streuung zu groß ist, auch wenn durch Aufnahme von Regressoren die konditionale Varianz noch verringert wird. Dies wird im Zählansatz bestätigt.

---

<sup>10</sup> Private Institutionen ohne Erwerbzweck, die in erster Linie für Unternehmen arbeiten (Wissenschaftsstatistik GmbH, 2001, S. 36).

Die berechnete Korrelation von 0,9841 läßt auf einen stark positiven Zusammenhang zwischen der Patentvariablen und den FuE-Beschäftigten schließen. Auf einen ähnlich hohen Zusammenhang weist die Korrelation von 0,9788 zwischen den Patentanmeldungen und den FuE-Ausgaben hin. Dies verdeutlicht die folgende Abbildung:

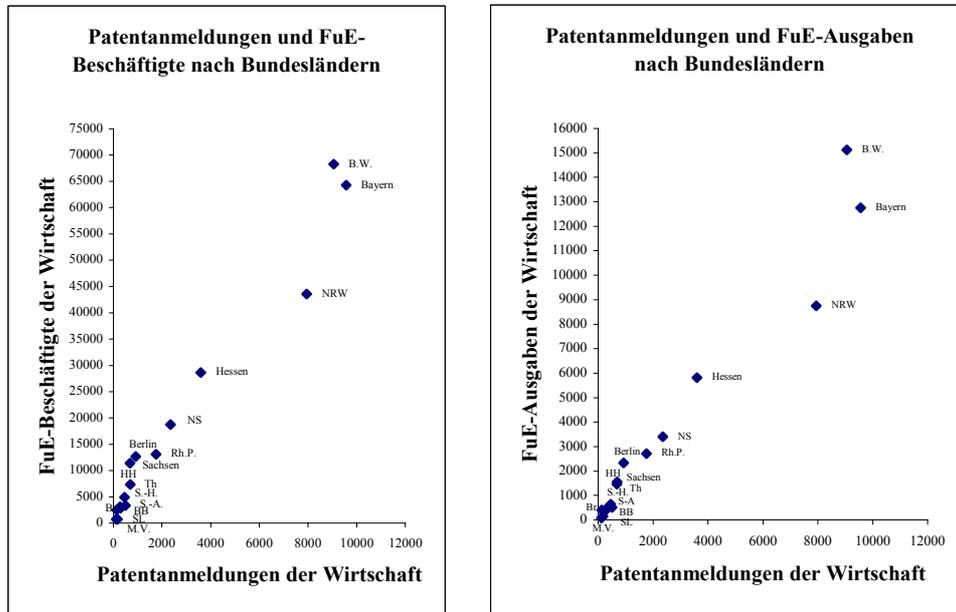


Abbildung: Patentanmeldungen und FuE-Input des Wirtschaftssektors nach BL

Die vermutete positive funktionale Abhängigkeit zwischen Patentanmeldungen und den FuE-Beschäftigtenzahlen bzw. den FuE-Ausgaben soll nun mit Zähldatenansätzen überprüft werden.

### 3.3 Spezifikation der Zähldatenansätze

Eine Vielzahl von empirischen Untersuchungen (z.B. Hausman/ Hall/ Griliches, 1984, S. 909-938 oder Meliciani, 2000, S. 1429-1437) wenden Zähldatenmodelle auf die Fragestellung an, wie die Zahl der Patentanmeldungen von FuE-Ausgaben abhängt. Dabei handelt es sich meist um Analysen im Paneldaten-Bereich. Ansätze für Querschnittsdaten sind seltener. Ein aktuelles Beispiel hierfür findet sich in Fritsch, 2002, S. 86-101. Hier werden die in Teil 2 theoretisch dargestellten Ansätze für Querschnittsdaten angewendet.

Die Anzahl der Patentanmeldungen bildet im folgenden die abhängige Zählvariable. Um die Werte für die Patente ganzzahlig zu machen, werden die Patentzahlen des Ausgangsdatensatzes gerundet. Da die Aufteilung auf Bundeslandebene erfolgt, liegen  $n = 16$  Beobachtungswerte vor. Gemäß der oben gefundenen Korrelationen werden zwei alternative Ansätze betrachtet.

In einem ersten Ansatz bildet die Zahl der FuE-Beschäftigten den Regressor. Hierbei liegt die Annahme zugrunde, daß Aufwendungen für FuE zu einem großen Teil aus Kosten für FuE-Beschäftigte bestehen (Fritsch, 2002, S. 91) und diese damit eine gute Proxyvariable für den Input in die FuE darstellen. Aufgrund der gefundenen positiven Korrelation ist zu erwarten, daß die Zahl der Patente mit zunehmender Beschäftigtenzahl ansteigt.

Die unabhängige Variable des zweiten Ansatzes bilden die FuE-Ausgaben (in Mio. DM). Diese Meßgröße beinhaltet neben den Personalkosten weitere interne Aufwendungen der Unternehmen für FuE und bildet daher den FuE-Input noch umfassender ab, als die Beschäftigtenzahlen. Mit zunehmenden Ausgaben wird auch hier mit einer steigenden Patentzahl gerechnet.

Der erwartete Wert der angemeldeten Patente bestimmt sich aus folgenden Gleichungen:

$$E(\text{Patente} \mid \text{Beschäftigte}) = \exp(\beta_{0B} + \beta_{1B} \text{Beschäftigte}) \text{ bzw.}$$

$$E(\text{Patente} \mid \text{Ausgaben}) = \exp(\beta_{0A} + \beta_{1A} \text{Ausgaben}).$$

### 3.4 Schätzung und Beurteilung der Zähldatenansätze

Die Parameter  $\beta$  werden jeweils für ein Poisson-Regressionsmodell und ein NegBin 2-Regressionsmodell mit Hilfe des Software-Pakets Limdep 7.0 (Greene, 1998, S. 595-628) bestimmt.

#### 3.4.1 Poisson-MLS

Nach Maximierung der Gleichung (5) erhält man folgende Regressionsergebnisse:

*Tabelle 3.1 Poisson-MLS-Ergebnisse*

Variable	Koeffizient $\beta_B$	Standardfehler $s_{\beta_B}$	Z-Wert $\beta_B / s_{\beta_B}$	Sign.level
<b>Konst.</b>	6,4432	0,010526	612,131	0,0000
<b>Beschäftigte</b>	0,0000430	0,000000199	215,764	0,0000

Variable	Koeffizient $\beta_A$	Standardfehler $s_{\beta_A}$	Z-Wert $\beta_A / s_{\beta_A}$	Sign.level
<b>Konst.</b>	6,5739	0,009865	666,420	0,0000
<b>Ausgaben</b>	0,0001924	0,000000887	216,961	0,0000

### Interpretation der Parameter

Der geschätzte Parameter  $\hat{\beta}_p$  einer unabhängigen Variablen gibt als Semi-Elastizität die relative (prozentuale) Änderung des konditionalen Mittelwertes an, die aus einer Änderung des Regressors um eine Einheit resultiert (Cameron/ Trivedi, 1996, S. 373; Winkelmann, 2000, S. 71).

Bei Veränderung der Beschäftigtenzahl um eine Einheit verändert sich die erwartete Patentzahl im Durchschnitt über alle Bundesländer daher um 0,00430 % (mit  $\hat{\beta}_{1Beschaef} = 0,000043$ ). Steigen die Ausgaben um eine Einheit (1 Mio. DM), erhöht sich die erwartete Patentzahl um 0,01924 % (mit  $\hat{\beta}_{1Ausgab} = 0,0001924$ ).

Die Absolutglieder geben Aufschluß darüber, welche Zahl von Patentanmeldungen zu erwarten ist, wenn die Regressorwerte alle gleich Null sind. Für beide Regressionen sind diese Werte positiv und liegen bei ca. 6 erwarteten Patenten. Dies können zum einen Patente sein, die ohne meßbaren FuE-Input entstehen, sogenannte „falling from heaven“-Innovationen (Fritsch, 2002, S. 90). Zum anderen können sie aus älterem, bestehenden Wissen des Unternehmens hervorgehen, welches mit den unabhängigen Variablen nicht darstellbar ist. Damit wäre die abhängige Variable falsch spezifiziert (Fritsch, 2002, S. 90).

Multipliziert man den Regressionsparameter mit dem Mittelwert der abhängigen Variablen, erhält man die durchschnittliche Erhöhung der Zählvariablen bei Änderung des Regressors  $j$  um eine Einheit (Cameron/ Trivedi, 1998, S. 80):

$$\frac{1}{n} \sum_{i=1}^n \frac{\partial E[Patente | Beschäft, Konst.]}{\partial Beschäftigte} = \hat{\beta}_{1B} \bar{y} = 0,000043 \cdot 2409,125 = 0,1036 ;$$

$$\frac{1}{n} \sum_{i=1}^n \frac{\partial E[Patente | Ausgab, Konst.]}{\partial Ausgaben} = \hat{\beta}_{1A} \bar{y} = 0,0001924 \cdot 2409,125 = 0,4635 .$$

So ist z.B. erwarten, daß bei einer Erhöhung der Beschäftigtenzahl um 1000 Personen, die durchschnittliche Zahl der angemeldeten Patente pro Bundesland um 103,6 ansteigt. Um ein Patent mehr zu entwickeln, muß ein Bundesland im Durchschnitt ca. 2,157 Mio. DM mehr in FuE investieren.

## Güte der Schätzung

Tabelle 3.2 Gütekriterien der Poisson-MLS

Unabh. Var.	<u>Beschäftigte</u>	<u>Ausgaben</u>
<b>Zahl d. Beob.werte</b>	16	16
<b>Vollst. Iterationen</b>	6	6
$\hat{L}_u$	- 4.135,92	-5.059,99
$\hat{L}_r$	-28.949,54	-28.949,54
$\chi^2$ -Wert	49.627,24	47.779,10
<b>Krit. Wert:</b> $\chi^2_{0,05;k=1}$	3,841	3,841
<b>Freiheitsgrade</b>	1	1
<b>Sign.level</b>	0,0000	0,0000
<b>Pearson-Residuum</b>	8.385,565	10.138,077
$R^2_{Pearson, Poisson}$	0,8795	0,8543
<b>Deviance-Residuum</b>	8.133,32	9.981,46
$R^2_{Deviance, Poisson}$	0,8592	0,8272

Die zwei Gütemaße zeigen bei beiden Regressionen einen „Erklärungsgehalt“ von über 80%. Nach dem  $R^2$ -Maß auf Basis der Deviance-Residuen, welches bei Hinzunahme von Regressoren nicht kleiner wird, hat die Poisson-Regression der Beschäftigtenzahl einen „Erklärungsgehalt“ von knapp 86%, während durch die Regression der Ausgaben etwa 83% „erklärt“ werden können. Beide Werte sind trotz der geringen Zahl von Beobachtungswerten recht hoch.

## Spezifikationstests

Die Größe der Stichprobenvarianz läßt bereits auf Overdispersion schließen. In einem solchen Fall wären die Standardfehlerschätzungen der Poisson-MLS verzerrt und die Signifikanztests ungültig. Deshalb wird zunächst auf Overdispersion auf Basis einer KQ-Regression getestet.

Nach Gleichung (27) und (28) wird mit den Regressoren

$$(29) \quad g(\hat{\mu}_i) = \hat{\mu}_i \sqrt{2} = w_{i1} ,$$

$$(30) \quad g(\hat{\mu}_i) = \hat{\mu}_i^2 \sqrt{2} = w_{i2}$$

folgende Regression geschätzt:

$$(31) \quad \frac{[(Patente_i - \hat{\mu}_i)^2 - Patente_i]}{\hat{\mu}_i \sqrt{2}} = \alpha g(\hat{\mu}_i) + \varepsilon_i .$$

Es ergibt sich:

Tabelle 3.3 Testergebnisse des Overdispersionstests mit KQ-Regression

Variable	Koeffizient ( $\alpha_B$ )	Stand.fehler ( $s_{\alpha_B}$ )	t-Wert ( $\alpha_B / s_{\alpha_B}$ )	Sign.level
$w_{i1\_Besch}$	523,2441	215,3745	2,429	0,0282
$w_{i2\_Besch}$	0,10202	0,05574	1,830	0,0871

Variable	Koeffizient ( $\alpha_A$ )	Stand.fehler ( $s_{\alpha_A}$ )	t-Wert ( $\alpha_A / s_{\alpha_A}$ )	Sign.level
$w_{i1\_Ausg}$	632,7864	259,0356	2,443	0,0274
$w_{i2\_Ausg}$	0,1473	0,0637	2,312	0,0354

Die t-Statistik einer Standard-KQ-Regression besitzt  $n - m - 1$  (hier =  $16 - 1 - 1 = 14$ ) Freiheitsgrade. Der kritische Wert ist daher  $t_{0,05;14} = 1,761$ . Da alle vier berechneten t-Werte den kritischen Wert überschreiten, kann  $H_0$  verworfen werden. Das bedeutet, daß alle  $\alpha$  signifikant von Null verschieden sind, und damit liegt in beiden Fällen Overdispersion vor. Signifikanztests der Poisson-Parameter wären somit ungültig.

Die berechneten Werte der LM-Statistik auf Overdispersion errechnen sich nach der Gleichung (24) zu:  $LM_{ber-Besch} = 1.416.047,17$  bzw.  $LM_{ber-Ausg} = 2.957.296,89$ . Der kritische Wert von 3,841 wird daher deutlich überschritten.  $H_0$  und damit die Equidispersion müssen auch hier in beiden Fällen abgelehnt werden.

Ein Grund für die festgestellte Overdispersion kann darin bestehen, daß die Anmeldung eines Patents weitere Patentanmeldungen im selben Zeitraum nach sich zieht. So können zum Beispiel neue Erkenntnisse über Krankheiten eine Vielzahl von Anmeldungen neuer Medikamentenpatente bewirken. Damit wäre die Unabhängigkeitsannahme verletzt. Um die Overdispersion zu berücksichtigen, wird nun ein NegBin 2-Modell geschätzt.

### 3.4.2 NegBin 2-MLS

Auf Basis der alternativen Spezifikation mit gammaverteilten Fehlertermen ermittelt man durch Maximierung der Gleichung (10) folgende Schätzwerte:

Tabelle 3.4 NegBin 2-MLS-Ergebnisse

Variable	Koeffizient $\beta_B$	Standardfehler $s_{\beta_B}$	Z-Wert $\beta_B / s_{\beta_B}$	Sign.level
<b>Konst.</b>	5,90664	0,212936	27,739	0,0000
<b>Beschäftigte</b>	0,0000612	0,00000746	8,209	0,0000
$\alpha$	0,3439	0,17704	1,943	0,0521

Variable	Koeffizient $\beta_{.A}$	Standardfehler $s_{\beta_{.A}}$	Z-Wert $\beta_{.A} / s_{\beta_{.A}}$	Sign.level
<b>Konst.</b>	6,00496	0,221813	27,072	0,0000
<b>Ausgaben</b>	0,0002871	0,00003628	7,912	0,0000
$\alpha$	0,3753	0,1857	2,021	0,0433

### Interpretation der Parameter

Die Interpretation der Parameter der NegBin 2-Schätzung unterscheidet sich nicht von der der Poisson-Parameter. Die Regressionskoeffizienten sind lediglich unterschiedlich groß.

Zur Angabe der Änderung des konditionalen Mittelwertes bei Änderung des Regressors um eine Einheit, wird wiederum der Regressionsparameter mit dem Mittelwert der abhängigen Variablen multipliziert. Hierbei ist zu beachten, daß bei der NegBin 2-Schätzung andere  $\mu$ 's bestimmt werden und sich damit die  $\bar{y}$  ändern. Mit  $\bar{y}_{Besch} = 3.584,63$  und  $\bar{y}_{Ausg} = 3.826,634$  folgt:

$$\frac{1}{n} \sum_{i=1}^n \frac{\partial E[Patente | Beschäft, Konst.]}{\partial Beschäftigte} = \hat{\beta}_{1B} \bar{y} = 0,0000612 \cdot 3.584,63 = 0,2194 ;$$

$$\frac{1}{n} \sum_{i=1}^n \frac{\partial E[Patente | Ausgab, Konst.]}{\partial Ausgaben} = \hat{\beta}_{1A} \bar{y} = 0,0002871 \cdot 3.826,63 = 1,0986 .$$

Es ist zu erwarten, daß bei Erhöhung der Beschäftigtenzahl um 1000 Personen die durchschnittliche Zahl der angemeldeten Patente pro Bundesland um 219,4 ansteigt. Um ein Patent mehr zu entwickeln, muß ein Land im Durchschnitt 0,91 Mio. DM mehr in FuE investieren.

### Güte der Schätzungen

Die  $R^2$ -Maße auf Basis der Deviance-Residuen in Tabelle 3.5 auf S. 19 liegen bei beiden Regressionen immer noch recht hoch. Die NegBin 2-Regression der Beschäftigtenzahl hat dabei einen „Erklärungsgehalt“ von etwa 82 %, während durch die Regression der Ausgaben ca. 79 % „erklärt“ werden können. Die  $R_{Deviance, NB}^2$  sind jedoch kleiner als die  $R_{Deviance, Poisson}^2$ . Dies kann für Daten mit zu großer Streuung auftreten, da die beiden  $R^2$ -Maße unterschiedliche Nenner haben und sie daher nicht direkt vergleichbar sind (Cameron/ Trivedi, 1998, S. 154).

Tabelle 3.5 Gütekriterien der NegBin 2-MLS

<b>Unabh. Var.</b>	<b><u>Beschäftigte</u></b>	<b><u>Ausgaben</u></b>
<b>Zahl d. Beob.werte</b>	16	16
<b>Vollst. Iterationen</b>	11	10
$\hat{L}_u = \text{NegBin 2}$	-124,224	-124,999
$\hat{L}_r = \text{Poisson}$	- 4.135,916	- 5.059,985
$\hat{L}_{Konst.}$	- 28.949,54	- 28.949,54
$\chi^2$ -Wert	8.023,384	9.869,973
<b>Krit.Wert</b> $\chi^2_{0,05;k=1}$	3,841	3,841
<b>Freiheitsgrade</b>	1	1
<b>Sign.level</b>	0,0000	0,0000
<b>Pearson-Residuum</b>	15,4721	14,7133
$R^2_{Pearson,NB}$	0,8156	0,8087
<b>Deviance-Residuum</b>	16,8947	16,9769
$R^2_{Deviance,NB}$	0,8119	0,7939

Nach der Gleichung (22) errechnen sich die LR-Statistiken auf Overdispersion der beiden Modelle zu:  $LR_{ber-Besch} = 8.023,384$  bzw.  $LR_{ber-Ausg} = 9.869,973$ . Der kritische Wert wird daher deutlich überschritten.  $H_0$  und damit die Equidispersion müssen auch hier in beiden Fällen abgelehnt werden. Die NegBin 2-Schätzung ist der Poisson-Schätzung vorzuziehen, da deren Verteilungsannahme den Daten besser gerecht wird.

Der einseitige, standardnormalverteilte Wald-Test auf Overdispersion der Koeffizienten  $\alpha$  auf Signifikanz ergibt bei einem kritischen Wert  $Z_{0,95} = 1,65$ , daß alle berechneten Z-Werte der Koeffizienten  $\hat{\alpha}_{NB2}$  größer sind als der kritische Wert, wenn auch nur gering.  $H_0$  kann also abgelehnt werden, alle  $\hat{\alpha}_{NB2}$  sind signifikant von Null verschieden, und man kann von zu großer Streuung ausgehen.

Die LR-Tests auf Signifikanz sind mit  $\chi^2_{Besch} = -2(-28.949,54 - (-124,224)) = 57.650,632$  und  $\chi^2_{Ausg} = -2(-28.949,54 - (-124,999)) = 57.649,082$  viel größer als der kritische Wert von 3,841. Die Nullhypothese, daß alle aufgenommenen Regressoren zusammen keinen Einfluß auf die Patentanmeldungen haben, kann somit verworfen werden.

Da die berechneten Werte des Wald-Tests auf Signifikanz aus der Tabelle 3.4 auf S. 17f alle größer sind als der kritische Wert von 1,96, können die Nullhypothesen abgelehnt werden. Damit besteht ein signifikanter Einfluß der beiden Konstanten sowie von Beschäftigten bzw. Ausgaben auf die Patentzahlen.

### 3.4.3 NegBin 2-QGPMLS

Um relativ effizientere Schätzwerte zu erhalten, wird eine NegBin 2-QGPMLS durch Maximierung der Gleichung (13) durchgeführt. Nach Formel (12) ergeben sich Schätzwerte für  $\alpha$  zu  $\tilde{\alpha}_{Besch} = 0,36913$  und  $\tilde{\alpha}_{Ausg} = 0,40027$  und damit folgende  $\hat{\beta}_{QGPML\_NB2}$ :

Tabelle 3.6 NegBin 2-QGPMLS-Ergebnisse

Variable	Koeffizient $\beta_B$	Standardfehler $s_{\beta_B}$	Z-Wert $\beta_B / s_{\beta_B}$	Sign.level
<b>Konst.</b>	5,906602	0,2276386	25,947	0,0000
<b>Beschäftigte</b>	0,0000612	0,000008005	7,648	0,0000

Variable	Koeffizient $\beta_A$	Standardfehler $s_{\beta_A}$	Z-Wert $\beta_A / s_{\beta_A}$	Sign.level
<b>Konst.</b>	6,00486	0,23516	25,535	0,0000
<b>Ausgaben</b>	0,0002871	0,000038482	7,460	0,0000

Auch hier sind alle berechneten Z-Werte des Wald-Tests größer als der kritische Wert von  $Z_{0,975} = 1,96$ . Damit bestätigt sich der signifikante Einfluß der Regressoren auf die Patentzahlen.

### 3.5 Vergleich der verschiedenen Schätzergebnisse

Die verschiedenen Schätzergebnisse sollen nun verglichen werden. Einen Überblick über alle Ergebnisse gibt *Tabelle 3.7* beispielhaft für den Regressor „Beschäftigte“:

Tabelle 3.7 Vergleich der Schätzergebnisse des Regressors „Beschäftigte“

Schätzmethode	Variable	Koeffizient $\beta_B$	Stand.fehler $s_{\beta_B}$	Z-Wert $\beta_B / s_{\beta_B}$
Poisson-MLS	<b>Konst.</b>	6,4432	0,01053	612,131
	<b>Beschäft.</b>	0,000043045	0,000000199	215,764
NB2-MLS	<b>Konst.</b>	5,9066	0,21294	27,739
	<b>Beschäft.</b>	0,000061220	0,000007458	8,209
NB2-QGPMLS	<b>Konst.</b>	5,9066	0,22764	25,947
	<b>Beschäft.</b>	0,000061226	0,000008005	7,648
	$\tilde{\alpha}$	0,36913	--	--

Die Poisson-ML-Schätzer sind nur dann konsistent, asymptotisch effizient und normalverteilt, wenn die spezifizierte Verteilung der wahren Verteilung entspricht. Da alle Overdispersionstests zu dem Schluß kommen, daß die Daten eine zu große Streuung aufweisen, sind die Poisson-ML-Schätzer nicht mehr effizient, und deren Standardfehler und Teststatistiken sind verzerrt. Deshalb wird ein NegBin 2-Modell geschätzt.

Wenn die Daten negativ binomialverteilt sind, ist die NegBin 2-MLS effizient mit korrekten Standardfehlern und bietet die optimale Schätzung. Liegen Verteilungsmaßspezifikationen vor, weist die NegBin 2-MLS falsche Standardfehler aus, und es sollte die QGPMLS verwendet werden. Da sich die geschätzten Werte der Koeffizienten für die NegBin 2-MLS und die NegBin 2-QGPMLS nur minimal unterscheiden, liegt kein Hinweis auf Verletzung der Annahme vor, daß die Zählvariable tatsächlich negativ binomialverteilt ist (Cameron/ Trivedi, 1986, S. 43). Zur Beurteilung des Zusammenhangs zwischen Patenten und FuE-Input sollte man demnach hier am zweckmäßigsten die NegBin 2-MLS-Ergebnisse verwenden.

Auch bei der Interpretation der Parameterschätzungen ist es wichtig zu wissen, ob ein NegBin- oder ein Poisson-Modell die Daten besser abbildet, da sich vor allem die partiellen Ableitungen der erwarteten Patentwerte in beiden Fällen stark unterscheiden. Sind nach der Poisson-Schätzung noch 2,157 Mio. DM an höheren FuE-Ausgaben erforderlich, um ein Patent mehr zu entwickeln, so sind es beim NegBin-Modell nur noch 0,91 Mio. DM. Die ökonomischen Schlußfolgerungen sind demnach sehr stark von der korrekten Spezifikation der Modelle abhängig.

#### **4 Zusammenfassung und Ausblick**

Aufgrund der festgestellten Signifikanz der Parameter der NegBin 2-MLS und den hohen  $R^2$ -Maßen bestätigt sich der in der Literatur dargestellte, stark positive Zusammenhang zwischen den Patentanmeldungen und den FuE-Beschäftigten bzw. den FuE-Ausgaben. Je mehr ein Unternehmen in FuE investiert, desto mehr Patente kann es anmelden und damit den technischen Fortschritt vorantreiben. Mit Hilfe eines geeigneten Managements kann es dadurch eine günstigere Wettbewerbsposition erreichen und den Unternehmenserfolg sichern. Firmen wie Siemens, Bosch oder DaimlerChrysler, die die Spitzenplätze in der Statistik der jährlichen Patentneuanmeldungen belegen (Deutsches Patent- und Markenamt, 2001, S. 17), schaffen sich durch ihre strategische Ausrichtung als innovative Unternehmen die Möglichkeit, neue Märkte und neue Nachfrage zu erschließen. Sie bewegen sich

oft außerhalb von bestehenden Branchenstandards und sind somit in vielen Bereichen unabhängig vom Wettbewerber.

In der Literatur werden zur Klärung des hier dargestellten Sachverhalts meist Paneldaten-Analysen durchgeführt, mit denen sich der Einfluß der zeitlichen Entwicklung von FuE-Ausgaben auf die Patentanmeldungen ermitteln läßt. Diese Ansätze stellen jedoch höhere Anforderungen an die Daten. Damit gehen auch Kostenfragen einher, da solche Daten meist originär erhoben werden müssen. Im Hinblick auf das Methodenwissen der Adressaten aus der Wirtschaft lassen sich die weniger komplexen Verfahren im Querschnittsdaten-Bereich leichter kommunizieren und interpretieren. Da im empirischen Teil der „Erklärungsgehalt“ des geschätzten NegBin 2-Modells hoch ist, können Handlungsoptionen für die Entscheidungsträger abgeleitet werden, auch ohne zeitliche Entwicklungen berücksichtigen zu müssen.

Mit weiterführenden Analysen von regionalen Aspekten und bundeslandspezifischen Effekten auf Patentanmeldungen könnten wichtige Erkenntnisse für die Unternehmen bspw. im Hinblick auf ihre Standortwahl gewonnen werden. Wie in der Abbildung auf S. 13 deutlich wird, unterscheiden sich einzelne Bundesländer stark im Grad ihrer Innovationsintensität. Es bleibt zu untersuchen, ob die spezifischen Eigenarten eines Bundeslandes Patentanmeldungen eher begünstigen.

Für den Nachweis einer solchen Abhängigkeit sind allerdings weitaus detailliertere Daten notwendig, als sie bei dieser Analyse zur Verfügung standen.

Tabelle 3.8 Datensatz

Nr.	Bundesland	Patentanmeldungen insgesamt 1998	Patentanmeldungen der Wirtschaft Anteil in % 1998	Patentanmeldungen der Wirtschaft 1998	Zahl der FuE Beschäft. 1997	Interne FuE Aufwendungen 1997 in Mio. DM
01	Schleswig-Holstein	656	0,763	501	3.340	512
02	Hamburg	893	0,780	697	7.359	1.549
03	Niedersachsen	2.966	0,793	2.352	18.763	3.405
04	Bremen	170	0,792	135	2.490	410
05	Nordrhein-Westfalen	9.657	0,822	7.938	43.568	8.754
06	Hessen	4.151	0,867	3.599	28.637	5.813
07	Rheinland-Pfalz	2.068	0,853	1.764	13.082	2.710
08	Baden-Württemberg	10.879	0,832	9.051	68.270	15.125
09	Bayern	11.691	0,818	9.563	64.288	12.747
10	Saarland	289	0,609	176	748	147
11	Berlin	1.405	0,663	932	12.708	2.334
12	Brandenburg	444	0,675	300	2.860	459
13	Mecklenburg-Vorpommern	215	0,531	114	724	79
14	Sachsen	1.036	0,657	681	11.438	1.456
15	Sachsen-Anhalt	415	0,684	284	3.073	398
16	Thüringen	698	0,657	459	4.922	642
	<b>BRD gesamt</b>	<b>47.633</b>	<b>0,807</b>	<b>38.546</b>	<b>286.270</b>	<b>56.540</b>

<b>Mittelwerte</b>	2.409,13	17.891,88	3.533,75
<b>Standardabw.</b>	3.342,96	22.090,77	4.704,76
<b>Varianzen</b>	11.175.360,78	488.002.297,1	22.134.750,33
<b>Korrelation</b>		0,984112207	0,978826514

## Literatur

### **Cameron, A.C./Trivedi, P.K.**

Econometric Models based on Count Data: Comparisons and Applications of some Estimators and Tests; *Journal of Applied Econometrics*, vol. 1, 1986, S. 29-53.

### **Cameron, A.C./Trivedi, P.K.**

Regression-Based Tests for Overdispersion in the Poisson Model  
*Journal of Econometrics*, vol. 46, 1990, S. 347-364.

### **Cameron, A.C./Trivedi, P.K.**

Count Data Models for Financial Data  
in: Maddala, G.S./Rao C.R. (eds.), *Handbook of Statistics – Statistical Methods in Finance*, vol. 14, Amsterdam usw. 1996, S. 363-391.

### **Cameron, A.C./Trivedi, P.K.**

*Regression Analysis of Count Data*; Cambridge University Press 1998.

### **Cameron, A.C./Windmeijer, F.A.G.**

R-Squared Measures for Count Data Regression Models With Applications to Health-Care Utilization; *Journal of Business & Economic Statistics*, vol. 14, 1996, S. 209-220.

### **Deutsches Patent- und Markenamt**

Jahresbericht 2000;  
<http://www.dpma.de/veroeffentlichungen/jahresbericht00/jb2000.pdf>  
Stand: 2001, Abruf: 24.01.2002, S. 1-73.

### **Dionne, G./Vanasse, C.**

Automobile Insurance Ratemaking in the Presence of Asymmetrical Information  
*Journal of Applied Econometrics*, vol. 7, 1992, S. 149-165.

### **Fritsch, M.**

Measuring the Quality of Regional Innovation Systems: A Knowledge Production Function Approach; *International Regional Science Review*, vol. 25, 2002, S. 86-101.

### **Greene, W.H.**

LIMDEP Version 7.0 - User's Manual, Revised Edition, Econometric Software Inc., Plainview N.Y., 1998.

### **Greene, W.H.**

*Econometric Analysis*, 4<sup>th</sup> ed., Prentice Hall, New York 2000.

### **Greif, S.**

Forschung und Entwicklung und Patente in: Herzog, R. (Hrsg.), *F&E-Management in der Pharma-Industrie*, Aulendorf 1995, S. 229-239.

### **Greif, S.**

Patentgeographie. Die räumliche Struktur der Erfindungstätigkeit in Deutschland  
*Raumforschung und Raumordnung*, Jg. 59, 2001, S. 142-153.

**Hausman, J./Hall, B.H./Griliches, Z.**

Econometric Models for Count Data with an Application to the Patents-R&D Relationship; *Econometrica*, vol. 52, 1984, S. 909-938.

**King, G.**

Variance Specification in Event Count Models: From Restrictive Assumptions to a Generalized Estimator; *American Journal of Political Science*, vol. 33, 1989, S.762-784.

**Meliciani, V.**

The relationship between R&D, investment and patents: a panel data analysis  
*Applied Economics*, vol. 32, 2000, S. 1429-1437.

**Ronning, G.**

Mikroökometrie; Springer-Verlag, Berlin usw. 1991.

**Winkelmann, R.**

Econometric Analysis of Count Data, 3<sup>rd</sup>, revised and enlarged edition,  
Springer-Verlag, Berlin usw. 2000.

**Winter, J.**

Ökonometrische Verfahren für Zähldaten und Verweildauern  
<http://www.sfb504.uni-mannheim.de/~joachim/lehre/01-SS/micro.htm>  
Stand: 10.03.2001, Abruf: 20.12.2001, S. 30-33.

**Wissenschaftsstatistik GmbH im Stifterverband für die Deutsche Wissenschaft  
(Hrsg.)**

FuE-Datenreport 2001 – Forschung und Entwicklung in der Wirtschaft  
1999-2000 – Bericht über die FuE-Erhebung 1999.

Autorin: cand. rer. pol. Verena Dexheimer, Projektbearbeiterin

Bisher erschienene Arbeitspapiere:

1. Peter M. Schulze, Prognoseverfahren wissenschaftlicher Institute in der Bundesrepublik Deutschland. Überblick über eine Umfrage (Dezember 1993)
2. Martina Nold / Peter M. Schulze, Möglichkeiten und Grenzen der Quantifizierung der Schattenwirtschaft (April 1994)
3. Armin Seher, Einfluß der Integrationsordnung bei Zeitreihen auf die Spezifikation von Fehlerkorrekturmodellen (Juni 1994)
4. Lars Berg / Armin Gemünden / Frank Hubert / Ralf Leonhardt / Michael Leroudier, Die Situation der Studentenschaft in den Wirtschaftswissenschaften an der Universität Mainz im Frühjahr 1994. Ergebnisse einer Umfrage (August 1994)
5. Christoph Balz, Ein Fehlerkorrekturmodell zur Entwicklung des Kapitalmarktzinses in der Bundesrepublik Deutschland (Oktober 1994)
6. Reinhard Elkmann / Nora Lauterbach / Stephan Wind, Tertiärisierung regionaler Wirtschaftsstrukturen. Eine empirische Analyse kreisfreier Städte und Landkreise in Hessen, Rheinland-Pfalz und dem Saarland (Dezember 1994)
7. Peter M. Schulze / Uwe Spieker, Deutsche Aktienindizes. Statistische Konzepte und Beispiele (Dezember 1994)
8. Armin Seher / Peter M. Schulze, Fehlerkorrekturmodelle und die Bewertung von Aktienkursindizes. Empirische Analyse zur Eignung des Konzepts (Januar 1995)
9. Reinhard Elkmann / Annette Klostermann / Kerstin Lieder, Zur intertemporalen Konstanz der Struktur regionaler Lohn- und Gehaltsniveaus in der Bundesrepublik Deutschland (Mai 1995)
10. Christoph Fischer, Ein Fehlerkorrekturmodell zur Kaufkraftparitätentheorie (März 1996)
11. Ralf Becker / Claudia Müller, Zur Schätzung regionaler Konsumfunktionen (Oktober 1996)
12. Frank Hubert, Klassifizierung der Arbeitsmärkte in den OECD-Ländern mittels Cluster- und Diskriminanzanalyse (April 1997)
13. Frank Hubert, Das Okun'sche Gesetz: Eine empirische Überprüfung für ausgewählte OECD-Länder unter besonderer Berücksichtigung der nationalen Arbeitsmarktordnungen (September 1997)
14. Christoph Balz / Peter M. Schulze, Die Rolle nationaler, regionaler und sektoraler Faktoren für die Variation von Output, Beschäftigung und Produktivität in der Bundesrepublik Deutschland (Dezember 1997)

15. Peter M. Schulze, Steigende Skalenerträge und regionales Wachstum: Eine quantitative Analyse mit kleinräumigen Daten (März 1998)
16. Ralf Becker, Die Verallgemeinerte Momentenmethode (Generalized Method of Moments - GMM). Darstellung und Anwendung (Juni 1998)
17. Peter M. Schulze, Regionales Wachstum: Sind die Dienstleistungen der Motor? (August 1998)
18. Ke Ma, Absatzanalyse für den chinesischen Pkw-Markt (Oktober 1998)
19. Christoph Balz/Peter M. Schulze, Die sektorale Dimension der Konvergenz. Eine empirische Untersuchung für die Bundesrepublik Deutschland (Januar 1999)
- 20.\* Robert Skarupke, Quantifizierung des Heimvorteils im deutschen Profifußball: Eine empirische Untersuchung für die 1. Fußball-Bundesliga (August 2000)
- 21.\* Peter M. Schulze, Regionalwirtschaftlicher Datenkatalog für die Bundesrepublik Deutschland (September 2000)
- 22.\* Yvonne Lange, Ein logistisches Regressionsmodell zur Analyse der Verkehrsmittelwahl im Raum Mainz (Oktober 2000)
- 23.\* Verena Dexheimer, Zählmodellen (Count Data Models) Ansätze und Anwendungen (Mai 2002)

---

\* Im Internet unter <http://wiwi.uni-mainz.de/vwl/schulze/welcome.html> verfügbar.